VARIABLES AND METHOD FOR AUTHORSHIP ATTRIBUTION

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

Some of the work in this application was supported by grants 95-IJ-CX-0012 and 98-LB-VX-0065 from the National Institute of Justice, Office of Justice Programs, United States Department of Justice. Points of view in this document are those of the author and do not represent the official position of the U.S. Department of Justice. The federal government may have an interest in this application.

CROSS-REFERENCES TO RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application Ser. No. 60/668,004, filed on 4 Apr. 2005, the contents of which are incorporated by reference herein in their entirety.

COPYRIGHT NOTICE

FIELD OF THE INVENTION

This invention relates to the field of determining the authorship of documents, by analyzing the structure of the language (i.e., the syntax, discourse and punctuation) used within the document. The method employed herein can be used to determine authorship of short textual works as well as more lengthy works such as a book, manuscript or the like, and can be utilized in a forensic setting.

BACKGROUND OF THE INVENTION

Introductory material is presented in this section, relating (A) specific principles guiding language-based authorship attribution within the forensic setting; (B) general principles of authorship attribution as a pattern-recognition problem; (C) background information in authorship attribution, including variables, methods and results of others, and (D) principles of syntax, markedness and part-of-speech tagging which underlay embodiments of the present invention.

A. Language-Based Authorship Attribution in the Forensic Setting.

During the course of criminal investigations, documents come to light whose authorship is uncertain but yet can be legally significant. Authorship determination is important in situations such as: a ransom note in a kidnaping; a threatening letter; anonymous letters; suicide notes; interrogation and/or interview statements; locating missing persons; employment disputes; examination fraud; plagiarism; will contests; peer review of reports in various other situations; and other contested issues of authorship. In view of the current focus on terrorism and the search for persons involved in terrorist acts, making terroristic threats, or kidnaping of citizens, the determination of authorship also plays a significant role.

While in the past these documents were generally handwritten, increasingly they are being produced with the aid of computers and printers, over electronic networks, or on printers or copiers, thus precluding the use of "standard" document analysis, which has typically focused on handwriting analysis, or analysis of the imprints of typewriter keys. In situations involving printed, electronically-produced or facsimile transmitted, rather than hand-written documents, the linguistic features of the document become important factors for determining the authorship of the document.

In contrast to handwriting examination or typewriter analysis, language-based authorship attribution relies on linguistic characteristics as variable sets for differentiating and identifying authors. In the literature on authorship attribution, there are four linguistic-variable classes which have been used by others and are sometimes combined with each other. These linguistic-variable classes are: (1) lexical, (2) stylometric, (3) graphemic, and (4) syntactic.

Lexical variables include vocabulary richness and function word frequencies; (function words in English are a closed set of words which specify grammatical functions, such as prepositions, determiners and pronouns).

Stylometric variables include word length, sentence length, paragraph length, counts of short words, and such.

Graphemic variables include the counts of letters and punctuation marks in a text.

Syntactic variables include the counts of syntactic part-of-speech tags such as noun, verb, etc., and adjacent part-of-speech tags.

As will be shown in the specification, and defined by the claims, new linguistic-variable sets are defined within these classes, and which variable sets are specifically applicable to authorship attribution in the forensic and non-forensic settings.

Authorship attribution in the forensic setting must meet certain criteria in order to be admitted as scientific evidence or entertained seriously as investigative support. In *Daubert v. Merrill-Dow Pharmaceuticals, Inc.,* 509 U.S. 579, 27 USPQ2d 1200 (1993), the Supreme Court set out guidelines which substantially changed the admissibility of scientific evidence within the federal court system, and which have become applicable in a number of state court jurisdictions as well. The criteria described herein are not those described in *Daubert*, but those that this inventor believes should guide the development of an authorship identification method, and which will later insure the admissibility of such evidence. Accordingly, these criteria are linguistic defensibility, forensic feasibility, statistical testability, and reliability.

First, the method must be linguistically defensible. Basic assumptions about language structure, language use, and psycholinguistic processing should undergird the method. The linguistic variables which are ultimately selected should be related in a straightforward way to linguistic theory and psycholinguistics; the linguistic variables should be justifiable. For example, function words have been used in many lexical approaches to authorship attribution, perhaps most famously by Mosteller and Wallace (1984). Function words can be justified as a potential discriminator for two reasons: first, function words are a lexical closed class, and second, function words are often indicators of syntactic structure. Psycholinguistically, function words are known as a distinct class for semantic processing and the syntactic structures which function words shadow are known to be real. A method based on function words is linguistically defensible